Занятие 5. Лекция

План

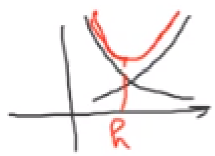
1. Возвращаем долги
2. Тестирование гипотез

Jackknife (=перочиный ножик)

- метод, позволяющий улучшить нахаляву оценку

На прошлом занятии обсуждали bias-variance-tradeof

$$MSE(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = \underbrace{Bias(\hat{\theta}_n)}_{=(\mathbb{E}\hat{\theta}_n - \theta)^2} + \underbrace{Var(\hat{\theta}_n)}_{=\mathbb{E}[(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2]}$$



- как уменьшить смещение?
  - jackknife
- как уменьшить разброс?
  - методы монте-карло

В теории хочется сделать так:

$$\hat{\theta}_n^0 = \hat{\theta}_n - \mathbb{E}\hat{\theta}_n d$$

Тогда получается $\mathbb{E}\hat{\theta}_n^0 = 0$, т.е. оценка несмещенная. Но как найти это матожидание?

Пример

$\theta = \sigma^2$: $\hat{\theta}_n = \frac{1}{n}\sum_{i=1} (X_i - \bar{X})^2$ – смещ. оц. д-ля θ

$\mathbb{E}\hat{\theta}_n = \frac{n-1}{n} \cdot \sigma^2$

Хочется вычесть, но вычесть не можем:

$\hat{\theta}_n^0 = \hat{\theta}_n - \frac{n-1}{n}\sigma^2$, но мы не знаем σ^2

Идея: вычесть два раза θ:

$$\hat{\theta}_n^0 = \hat{\theta}_n - Bias(\hat{\theta}_n)$$

$$\hat{\theta}_n^0 - \theta = \hat{\theta}_n - \theta - Bias(\hat{\theta}_n) \quad \Big| \, \mathbb{E}[\cdot]$$

$$Bias(\theta_n^1) = Bias(\hat{\theta}_n) - Bias(\hat{\theta}_n) = 0$$

Значит θ_n^1 несмещенная. Но мы не знаем $Bias(\hat{\theta}_n)$. Поэтому берем оценку (она называется jackknife):

$$\widehat{Bias(\hat{\theta}_n)} = (n-1)\Big(\frac{1}{n}\sum_{i}^{n}\hat{\theta}_{(-i)} - \hat{\theta}_n\Big)$$

Тогда $\mathbb{E}[\widehat{Bias(\hat{\theta}_n)}] = Bias(\hat{\theta}_n) + O(\frac{1}{n^2})$

Пример

$\theta = \sigma^2$; $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}(X_i - \bar{X})^2$

$\mathbb{E}\hat{\theta}_n = \frac{n-1}{n}\sigma^2 \Rightarrow bias(\hat{\theta}_n) = \mathbb{E}\hat{\theta}_n - \sigma^2 = -\frac{\sigma^2}{n}$

Метод jacknife:

$$\mathbb{E}[\widehat{bias}(\hat{\theta}_n)] = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \theta) - (\hat{\theta}_n - \theta)\right)$$

$$= (n-1)\left(bias(\hat{\theta}_{(-i)}) - bias(\hat{\theta}_n)\right)$$

$$= (n-1)\left(-\frac{\sigma^2}{n-1} + \frac{\sigma^2}{n}\right) = \underbrace{-\frac{\sigma^2}{n}}_{\text{оценка несмещ.}} = bias(\hat{\theta}_n), \text{ т.е.}$$

Получаем несмещенную оценку!!!

Типовая ситуация

Есть члены порядка $1/n, 1/n^2$

$$Bias(\hat{\theta}_n) = \frac{a}{n} + \frac{b}{n^2} + \dots$$

Тогда

$$\mathbb{E}[\widehat{bias}(\hat{\theta}_n)] \overset{(2)}{=} (n-1)\left(\frac{a}{n-1} + \frac{b}{(n-1)^2} - \frac{a}{n} - \frac{b}{n^2} + \dots\right) = \frac{a}{n} + O\left(\frac{1}{n^2}\right)$$

В итоге получаем:

$$\hat{\theta}_n^0 := \hat{\theta}_n - \widehat{bias}(\hat{\theta}_n)$$

$$bias(\hat{\theta}_n^0) = \mathbb{E}\hat{\theta}_n - \mathbb{E}[\widehat{bias}(\hat{\theta}_n)] - \theta = bias(\hat{\theta}_n) - \mathbb{E}[\widehat{bias}(\hat{\theta}_n)] =$$

$$= -\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = 0$$

В случае типовой ситуации получаем

$$bias(\hat{\theta}_n^0) = O(\frac{1}{n^2})$$

Статистические тесты

- основной инструмент в матстатистике, чтобы доказать факт по данным

Основные вещи

Есть выборка взятая из закона распределения:

$$X_1, \dots X_n \sim \mathcal{P}_\theta$$

Проверяем гипотезу:

$$\mathcal{H}_0 : \theta = \theta_0$$

Нужно найти C - критическую область, т.ч.

$$\mathbb{P}_{\theta_0}\{(X_0, \dots X_n) \in C\} = \alpha = 0.05$$

Если мы попали в это множество, то отвергаем гипотезу

Пример (Нераскрытые парашуты)

Тестируем парашют с номером i:

$$X_1, \dots X_n = \begin{cases} 1, & \text{парашют раскрылся}, & p = \theta \\ 0, & \text{иначе}, & p = 1 - \theta \end{cases}$$

Гипотеза:

$$\mathcal{H}_0 : \theta = 0.0001$$

Критическое множество:

$$C = \{X_1 + \dots + X_n \geqslant t_\alpha\}$$

(число нераскрытых парашутов больше альфа).

$$\mathbb{P}_{\theta_0}\{X_1 + \dots X_n \geqslant t_\alpha\} = \alpha$$
$$= \sum_{k=\lceil t_\alpha \rceil}^{n} C_n^k \theta_0^k (1 - \theta_0)^{n-K} = \alpha$$

Нужно решить это уравнения относительно α. Можно с помощью R. Но есть другой способ, спомощью ЦПТ:

$$\mathbb{P}_{\theta_0}\{X_1 + \dots X_n \geqslant t_\alpha\} = \mathbb{P}_{\theta_0}\Bigg\{\underbrace{\frac{X_1 + \dots X_n - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}}_{\sim N(0,1)} \geqslant \frac{t_\alpha - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\Bigg\} = \alpha$$

Значит,

$$\frac{t_\alpha - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = z_{1-\alpha} \;\Rightarrow\; t_\alpha = n\theta_0 + z_{1-\alpha}\sqrt{n\theta_0(1-\theta_0)}$$

Мы считаем количество раз, сколько раскрылся парашют, с числом t_α. Если больше, то отклоняем. Иначе **не отклоняется**, но не принимается.

Но тогда что мы принимаем?

Число a, которые было выше, называется ошибкой 1 рода. Еще есть ошибки второго рода

Ошибки второго рода

$$\mathbb{P}_{\theta_0}\{(X_0, \dots X_n) \in C\} = \alpha = 0.05$$

Идеально:

$$\mathbb{P}_{\theta_1}\{(X_1, \dots, X_n) \in C\} = 1 - \beta$$

где β - малое число. Оно является вероятностью ошибки второго рода.

А может ли быть так, что и $\alpha = \beta = 0$?

Пример

$X_1, \dots X_n$ H_0: норм. расп

H_0: расп Пуассона

$$\mathcal{P}_\Theta = \{\underset{\text{норм}}{\mathbb{P}_0}, \underset{\text{пуасс}}{\mathbb{P}_1}\}$$

Для этой задачи есть тест, для которого оба числа равны нулю

$$C = \{X_1 \in \mathbb{N} \cup \{0\}\}$$

Тогда:

$$\mathbb{P}_0\{X_1 = 0, 1, 2, \dots\} = \sum_{k=0}^{\infty} \mathbb{P}\{X_1 = k\} = 0 \Rightarrow \alpha = 0$$

$$\mathbb{P}_1\{X_1 = 0, 1, 2, \dots\} = 1 \Rightarrow \beta = 0.$$

Но это математическая экзотика

Какой размер выборки брать для эксперимента?

Максимально прикладной вопрос (вспоминаем тинек). Можно решить с помощью чисел α, β.

Пример про парашюты (продолжение)

$$P_{\theta_0}\left\{\frac{X_1+\dots+X_n-n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \geq \frac{t_\alpha-n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\right\} = \alpha$$

$$\approx \xi \sim N(0,1)$$

$$P_{\theta_1}\left\{\frac{X_1+\dots+X_n-n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} \geq \frac{t_\alpha-n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}}\right\} = 1-\beta$$

Мы ранее вывели

$$\begin{cases} \dfrac{t_\alpha-n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = z_{1-\alpha} \\ \dfrac{t_\alpha-n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} = z_\beta \end{cases}$$

Откуда (приравниваем t_α)

$$n\theta_0 + \sqrt{n\theta_0(1-\theta_0)}\, z_{1-\alpha} = n\theta_1 + \sqrt{n\theta_1(1-\theta_1)}\, z_\beta$$

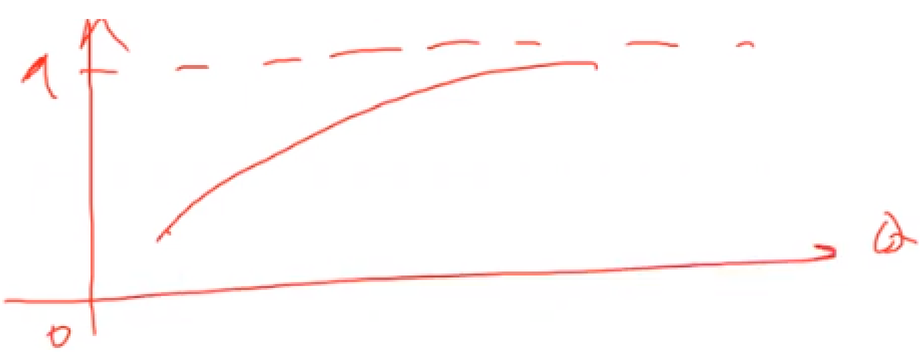
$$n = \left(\frac{\sqrt{\theta_1(1-\theta_1)}\, z_\beta - \sqrt{\theta_0(1-\theta_0)}\, z_{1-\alpha}}{\theta_0-\theta_1}\right)^2$$

Графическая иллюстрация ошибок первого и второго рода

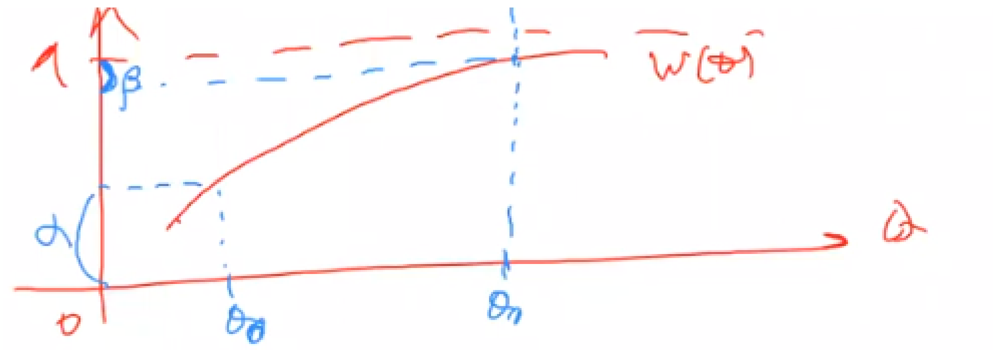
Функция мощности:

$$W(\theta) = P_\theta\{(x_1,\dots,x_n) \in C\}$$

Типичный график:

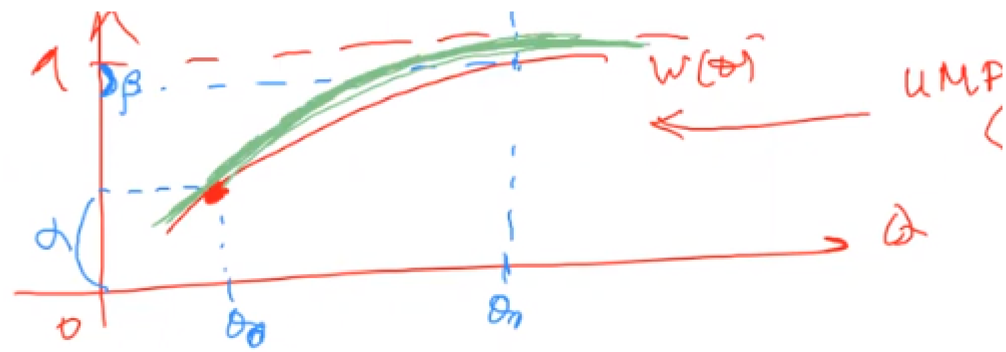


Вот так выглядят ошибки первого и второго рода:



- по параметру α понять как делать тест
- по параметру β можем понять длякаких гипотез метод эфективный

Можно придумать тест, у которого α фиксирована, а кривая оптимальна? Да, это называется UMP tests (uniformly most powerful)



Такие тесты можно строить благодаря теореме Неймана Пирсона

UMP тесты

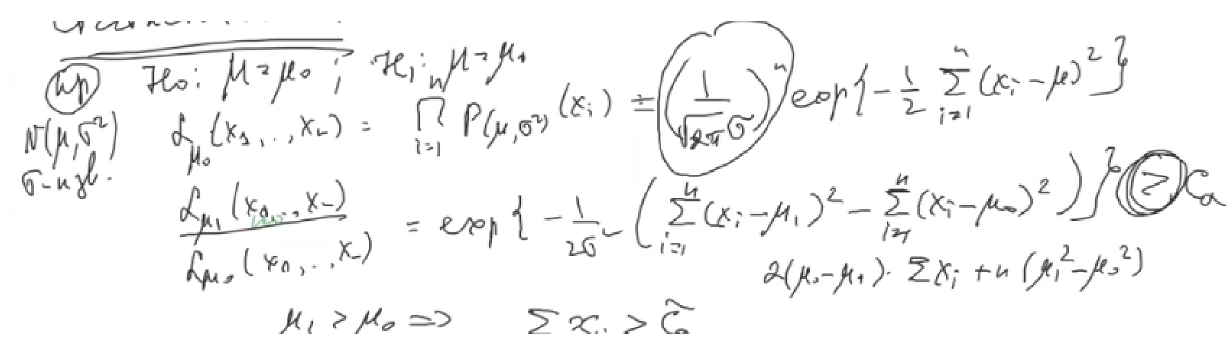
Теорема Неймана Пирсона

Есть две гипотезы $\theta = \theta_0$ и $\theta = \theta_1$. Тогда следующее критическое множество дает UMP тест:

$$C = \left\{ \frac{L_{\theta_1}(X_1, \dots X_n)}{L_{\theta_0}(X_1, \dots X_n)} > C_\alpha \right\}$$

где L - функция правдоподобия

Пример



Методика построения теста:

$$\underbrace{\sum X_i}_{\sim N(\mu_0 n,\, n^2\sigma^2)}$$

$$\mathbb{P}_{\mu_0}\{\sum X_i > \hat{c}_\alpha\} = \alpha$$

Тогда

$$\xi \sim N(0,1) \Rightarrow \{\sum X_i > \hat{c}_\alpha\} = \{\mu_0 n + n\sigma\xi > \hat{c}_\alpha\}$$

$$\{\xi > \underbrace{\frac{\hat{c}_\alpha - \mu_0 n}{n\sigma}}_{= z_\alpha}\}$$

(Нейман–Пирсон)

И в итоге

$$\hat{c}_\alpha = \mu_0 n + n\sigma z_\alpha$$

$\sum X_i > \hat{c}_\alpha$ — откл. гип.

$\sum X_i \le \hat{c}_\alpha$ — не откл. гип.

И опять экспоненциальное распределение

Оно идеально подходит под теорию.

Мы свели неравенство из теоремы НП к неравенству с суммоу X_i . В случае эксп. распр. так будет всегда

Экспоненциальное семейство:

$$p(x,\theta) = g(x)e^{x\theta - d(\theta)}$$

Посчитаем:

Статистич. тесты

Эксп. семейство: $p(x,\theta) = g(x)e^{x\theta - d(\theta)}$

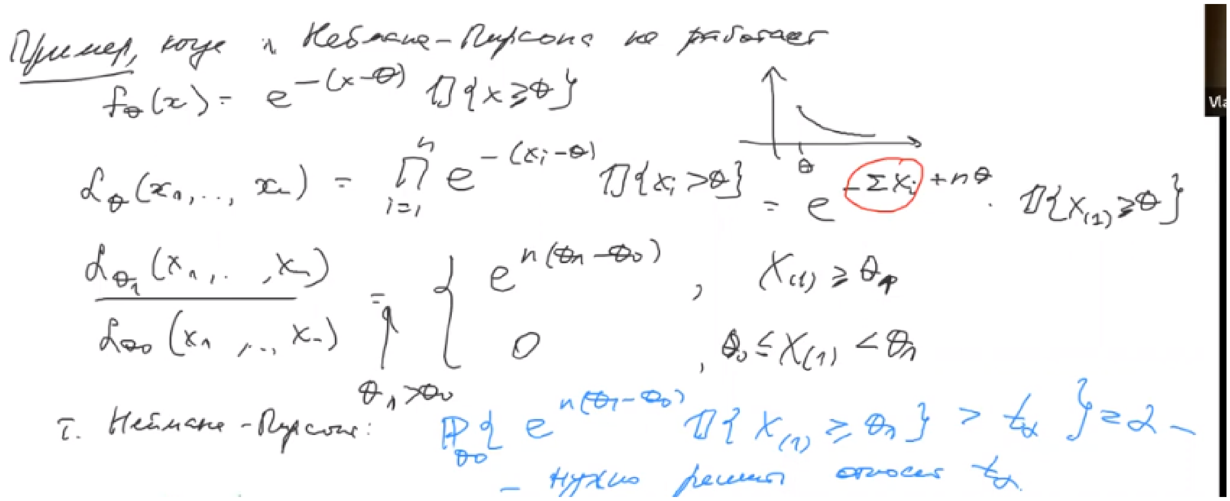
$$L_\theta(x_1,\dots,x_n) = g(x_1)\cdots g(x_n)e^{\sum_{i=1}^n x_i\theta - n d(\theta)}$$

$$\frac{L_{\theta_1}(x_1,\dots,x_n)}{L_{\theta_0}(x_1,\dots,x_n)} = e^{\sum x_i(\theta_1-\theta_0) - n(d(\theta_1) - d(\theta_0))} > c_\alpha$$

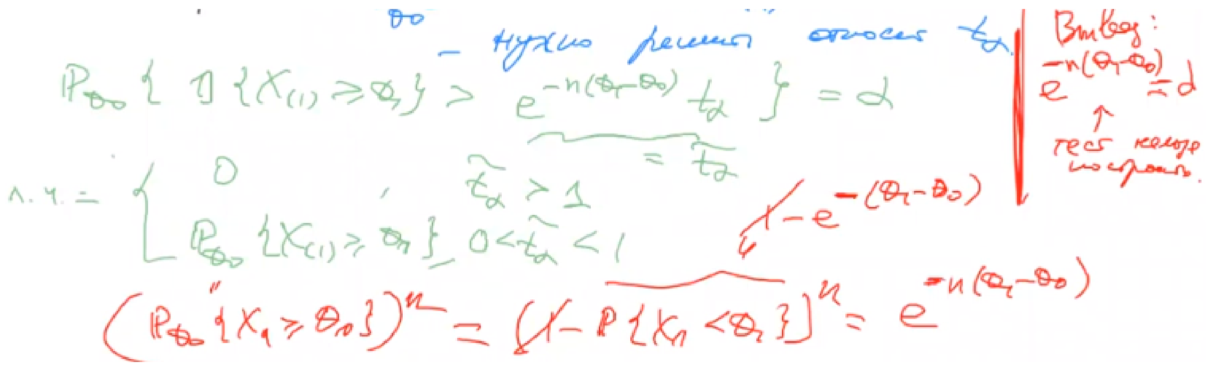
$$\theta_1 > \theta_0 \Rightarrow \boxed{\sum_{i=1}^n x_i > \hat{c}_\alpha}$$

Метод постр. UMP test: fix α; $\mathbb{P}_{\theta_0}\{\sum X_i > \hat{c}_\alpha\} = \alpha$ — для отыск. $\hat{c}$

Пример, когда лемма Немлана-Пирсона не работает



Последнее уравнение решить невозможно:



Получили в конце странное равенство $\Rightarrow$ тест построить не удается

Позже обсудим, что делать

Рандомизированный статистический тест

- не является тесто v в строгом смысле этого слова

Обычный тест:

$$\mathbb{P}_{\theta_0}\{(X_1, \ldots X_n) \in C\} = \alpha$$
$$\Leftrightarrow \mathbb{E}[\mathbb{I}\{(X_1, \ldots X_n) \in C\}]$$

Рандомизированный тест: вводим функцию d:

$$d(X_1, \ldots X_n) \in [0, 1]$$
$$\mathbb{E}_\infty[d(X_1, \ldots X_n)] = \alpha$$

Интерпретация:

- $d(X_1, \ldots X_n) = 1$ - отклоняем
- $d(X_1, \ldots X_n) = 0$ - не отклоняем
- $d(X_1, \ldots X_n) \in (0, 1)$ - отклоняем с вероятностью $d(X_1, \ldots X_n)$

Аналог леммы Неймана-Пирсона:

$$d(X_1, \dots X_n) = \begin{cases} 1, & \frac{L_{\theta_0}(X_1,\dots,X_n)}{L_{\theta_1}(X_1,\dots,X_n)} > t_\alpha \\ \gamma, & (\dots) = t_\alpha \\ 0, & (\dots) < t_\alpha \end{cases}$$

Тогда

1. $\exists! \gamma, t_\alpha: \ \mathbb{E}_{\theta_0}[d(X_1, \dots X_n)] = \alpha$
2. Такой тест самый мощный среди всех тестов с вероятностью ошибки первого рода α, т.е.

$$\forall d^* : \mathbb{E}_{\theta_0}[d^*(X_1, \dots X_n)] = \alpha \mid \mathbb{E}_\theta[d^*(X_1, \dots X_n)] \underset{\forall \theta}{\leqslant} \mathbb{E}_\theta[d(X_1, \dots X_n)]$$

Пример

Рандомизир. тест.

$$f_\theta(x) = e^{-(x-\theta)} \mathbb{1}\{x > \theta\}$$

$$\frac{L_{\theta_1}(x_1, \dots, x_n)}{L_{\theta_0}(x_1, \dots, x_n)} = e^{n(\theta_1 - \theta_0)} \cdot \mathbb{1}\{X_{(1)} > \theta_1\}$$

$$e^{n(\theta_1 - \theta_0)} \mathbb{1}\{X_{(1)} > \theta_1\} > t_\alpha$$

$$\mathbb{1}\{X_{(1)} > \theta_1\} > t_\alpha \cdot e^{-n(\theta_1 - \theta_0)} =: C_\alpha$$

$$d(X_1, \dots, X_n) = \begin{cases} 1, & \mathbb{1}\{X_{(1)} > \theta_1\} > C_\alpha \\ \gamma, & \mathbb{1}\{X_{(1)} > \theta_1\} = C_\alpha \\ 0, & \mathbb{1}\{X_{(1)} > \theta_1\} < C_\alpha \end{cases} \iff C_\alpha = 0 \text{ или } C_\alpha = 1$$

$C_\alpha = 0$: $\mathbb{E}_{\theta_0}[d(X_1, \dots, X_n)] = \mathbb{E}[\mathbb{1}\{X_{(1)} > \theta_1\} + \gamma \mathbb{1}\{X_{(1)} \leq \theta_1\}]$
$= \mathbb{P}\{X_{(1)} > \theta_1\} + \gamma \mathbb{P}\{X_{(1)} \leq \theta_1\} = e^{-n(\theta_1 - \theta_0)} + \gamma(1 - e^{-n(\theta_1 - \theta_0)}) = \alpha$

$$\gamma = \frac{\alpha - e^{-n(\theta_1 - \theta_0)}}{1 - e^{-n(\theta_1 - \theta_0)}} \in [0, 1] \iff e^{-n(\theta_1 - \theta_0)} < \alpha$$

$C_\alpha = 1$: $\mathbb{E}_{\theta_0}[d(X_1, \dots, X_n)] = \mathbb{E}_{\theta_0}[\gamma \cdot \mathbb{1}\{X_{(1)} \leq \theta_1\}] =$
$= \gamma \cdot P\{X_{(1)} \leq \theta_1\} = \gamma e^{-n(\theta_1 - \theta_0)} = \alpha$

$$\gamma = \alpha e^{n(\theta_1 - \theta_0)} \in [0, 1] \iff e^{-n(\theta_1 - \theta_0)} \geqslant \alpha$$

Когда $(\dots) = \alpha$, то используем обычную лемму НП

LR-тесты

У нас были очень простые гипотезы: параметр равен одному числу или другому. Обычно тесты другие: параметр равен заданному числу или нет:

$$\mathcal{H}_0 : \ \theta \in \Theta_0 \subset \Theta$$
$$\mathcal{H}_1 : \ \theta \in \Theta_1 \subset \Theta$$

Очень хочется обобщить лемму НП:

$$\left\{ \frac{\max_{\theta\in\Theta_1} L_\theta(x_1,\dots,x_n)}{\max_{\theta\in\Theta} L_\theta(x_1,\dots,x_n)} > C \right\}$$

Класс таких тестов называется LR.

Самый важные результат в этой области:

Теорема Уилкса

Пусть $\Theta_0 \subset \Theta$. Тогда

$$2\log\left(\frac{\max_{\theta\in\Theta} L_\theta(X_1,\dots X_n)}{\max_{\theta\in\Theta_0} L_\theta(X_1,\dots X_n)}\right) \xrightarrow{d} \mathcal{X}^2_{\dim\Theta-\dim\Theta_0}$$

где $X_p^2 = \xi_1^2 + \xi_p^2,\ \xi_i \sim N(0,1)$

У этой теоремы есть предположения, но опустим их (это условия регулярности). Они почти всегда выполнены.

Пример

Пр: $X_1,\dots,X_n \sim N(\mu,\sigma^2)$, σ^2 – изв $\quad H_0: \mu=\mu_0 \quad H_1: \mu\neq\mu_0 \qquad \xi_1,\dots\xi_p \sim N(0,1)$

$$L_\mu(x_1,\dots,x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}$$

$$\underbrace{\log\left(\max_\mu L_\mu(x_1,\dots,x_n)\right)} = \max_\mu\left(-n\log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right)$$

$$= -n\log(\sqrt{2\pi}\sigma) - \frac{\sum (x_i-\bar{x})^2}{2\sigma^2}$$

$$\log\left(\max L_\mu(x_1,\dots,x_n)\right) = -n\log(\sqrt{2\pi}\sigma) - \frac{\sum (x_i-\mu_0)^2}{2\sigma^2}$$

$$2\log(\dots) = \frac{n}{\sigma^2}(\bar{x}-\mu_0)^2 \xrightarrow{d} \chi^2_1$$

В нашем примере тут вообще равенство, а не стремление

Теорема Пирсона (1900-ый год)

Теорема Пирсона $\quad X_1,\dots,X_n$ – i.i.d. $\qquad \xi_1,\dots\xi \sim N(0,1)$

$D_1, D_2, D_3, \dots, D_k \qquad \bigsqcup_{i=1}^k D_i = \mathbb{R} \qquad P\{X\in D_i\} = q_i,\ i=1..k$

$N_i = \{$кол-во элем. выборки $\in D_i\}$

Утв: $\boxed{\sum_{i=1}^k \frac{(N_i - nq_i)^2}{nq_i} \xrightarrow[n\to\infty]{d} \chi^2_{k-1}}$

Прим: $X \sim F \leftarrow$ fixed $\quad q_i = P\{X\in D_i\} = F(u_{i+1}) - F(u_i)$, χ^2_{k-1}

$D_i = [u_i, u_{i+1}]$

α: $P\{\eta > t_\alpha\} = 1-\alpha$

$\sum\frac{(N_i-nq_i)^2}{nq_i} > t_\alpha$ – гип. отк.

Трудности: 1) $\xrightarrow{d}$.

Ивченко, Медведева (1984): $\begin{cases} nq_i \geq 5 \ \forall i \\ n \geq 50 \end{cases}$

2) $F \to F_\theta$ с неизв θ

Решение первой проблемы - эмпирические улосвия. Решение второй проблемы (через матан):

метод минимума χ^2: $q_i = F(a_{i+1}) - F(a_i) = q_i(\theta)$

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{k} \frac{(N_i - nq_i(\theta))^2}{nq_i(\theta)} \Rightarrow \sum_{i=1}^{k} \frac{(N_i - nq_i(\hat{\theta}_n))^2}{nq_i(\hat{\theta}_n)} \xrightarrow[n\to\infty]{d} \chi^2_{k-1-\dim\Theta}$$

Почему из теоремы Уилкса следует теорема Пирсона?

- просто полезная техника, характерная для этой области матанализа

Теорема Уилкса (Wilks thm)

$\Theta_0 \subset \Theta$

$$2\log\left(\frac{\max_{\theta\in\Theta} L_\theta(x_1,\dots,x_n)}{\max_{\theta\in\Theta_0} L_\theta(x_1,\dots,x_n)}\right) \xrightarrow{d} \chi^2_{\dim\Theta - \dim\Theta_0}$$

$\chi^2_p : \xi_1^2 + \dots + \xi_p^2$, $\xi_1, \dots, \xi_p \sim N(0,1)$

Теорема Пирсона

$X_1, \dots, X_n$ — i.i.d.

$D_1, D_2, D_3, \dots, D_k$; $\bigsqcup_{i=1}^{k} D_i = \mathbb{R}$; $P\{X \in D_i\} = q_i$, $i = 1..k$

$N_i = \{$кол-во элем. выборки $\in D_i\}$

Утв: $$\boxed{\sum_{i=1}^{k} \frac{(N_i - nq_i)^2}{nq_i} \xrightarrow[n\to\infty]{d} \chi^2_{k-1}}$$

Док. $P\{N_1 = n_1, \dots, N_k = n_k\} = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}$ — мультином. расп.

Θ_0: $p_i = q_i$, $\dim\Theta_0 = 0$

Θ: $p_1 + \dots + p_k = 1$, $\dim\Theta = k-1$

$$\log(\max L_\theta(x_1,\dots,x_n)) = \frac{n!}{n_1! \dots n_k!} \left(\frac{n_1}{n}\right)^{n_1} \dots \left(\frac{n_k}{n}\right)^{n_k}$$

В конце использовали

$$\frac{\partial}{\partial p_i}\left(n_1 \log p_1 + \dots + n_k \log p_k - \lambda(p_1 + \dots + p_k - 1)\right) = \frac{n_i}{p_i} - \lambda \Rightarrow p_i = \frac{n_i}{\lambda}$$

$\lambda = n$

Ок, идем дальше. Подставляем в формулу из теоремы Уилкса:

$$2\log\left(\frac{\left(\frac{n_1}{n}\right)^{n_1} \cdots \left(\frac{n_k}{n}\right)^{n_k}}{q_1^{n_1} \cdots q_k^{n_k}}\right) = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{nq_i}\right)$$

Разложим в ряд Тейлора:

Ф. Тейлора: $$x\log\left(\frac{x}{x_0}\right) = (x-x_0)\left(x\log\frac{x}{x_0}\right)' + \frac{1}{2}(x-x_0)^2\left(x\log\frac{x}{x_0}\right)'' + \dots = (x-x_0) + \frac{(x-x_0)^2}{2x_0} + \dots$$

$$= \cancel{2}\Bigg(\underbrace{\sum_{i=1}^{k} (n_i - nq_i}_{\sum n_i - n\sum q_i = 0,\ \sum n_i = n,\ \sum q_i = 1}) + \frac{(n_i - nq_i)^2}{\cancel{2}nq_i}\Bigg) = \sum \frac{(n_i - nq_i)^2}{nq_i} \quad \square$$

С теоремой Пирсона будет работать на семинарах.

Критерий хи-кварат для таблиц сопряженности

Таблица сопряженности (contingency table):

результ. обучения \ возраст	25-30	31-35	...
удовл			
неудовл.			

Можно было бы использовать корреляцию Пирсона. Но это предпологает, что распределение величин нормальное

Критерий хи-квадрат для таблиц сопряж.

contingency table

результ. обучения \ возраст	25-30	31-35	...
удовл	.	.	.
неудовл.	.	.	.

} и строки

и столбцов

$A_i = \{$ случ. вел X_i обладает св-вом из i-ой строки $\}$

$B_j = \{$ " — " j-ого столбца $\}$

$\Theta = \{ p_{ij},\ i = 1..n,\ j = 1..m \}$ $\qquad \dim \Theta = n \cdot m - 1$

$p_{ij} = P\{X \in A_i \cap B_j\}$ $\qquad \sum_{i,j} p_{ij} = 1$

$A_i \perp\!\!\!\perp B_j \quad \forall i, j$

$p_{ij} = P\{X \in A_i\} P\{X \in B_j\}$

$\dim \Theta_0 = \underbrace{n-1}_{P\{X \in A_i\}} + \underbrace{m-1}_{P\{X \in B_j\}}$

Т. Уилкса:

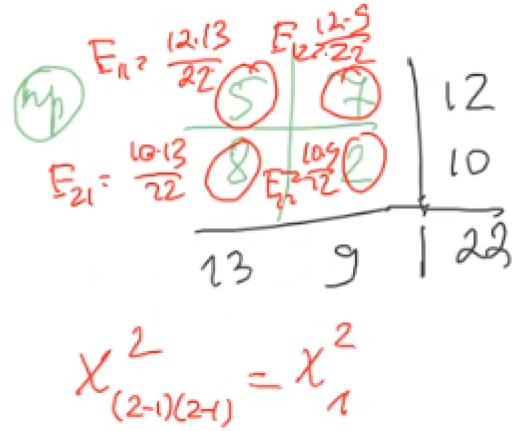
$\dim \Theta - \dim \Theta_0 = nm - 1 - (n + m - 2) = (n-1)(m-1)$

Непосредственно сам критерий:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(N_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{d} \chi^2_{(n-1)(m-1)}$$

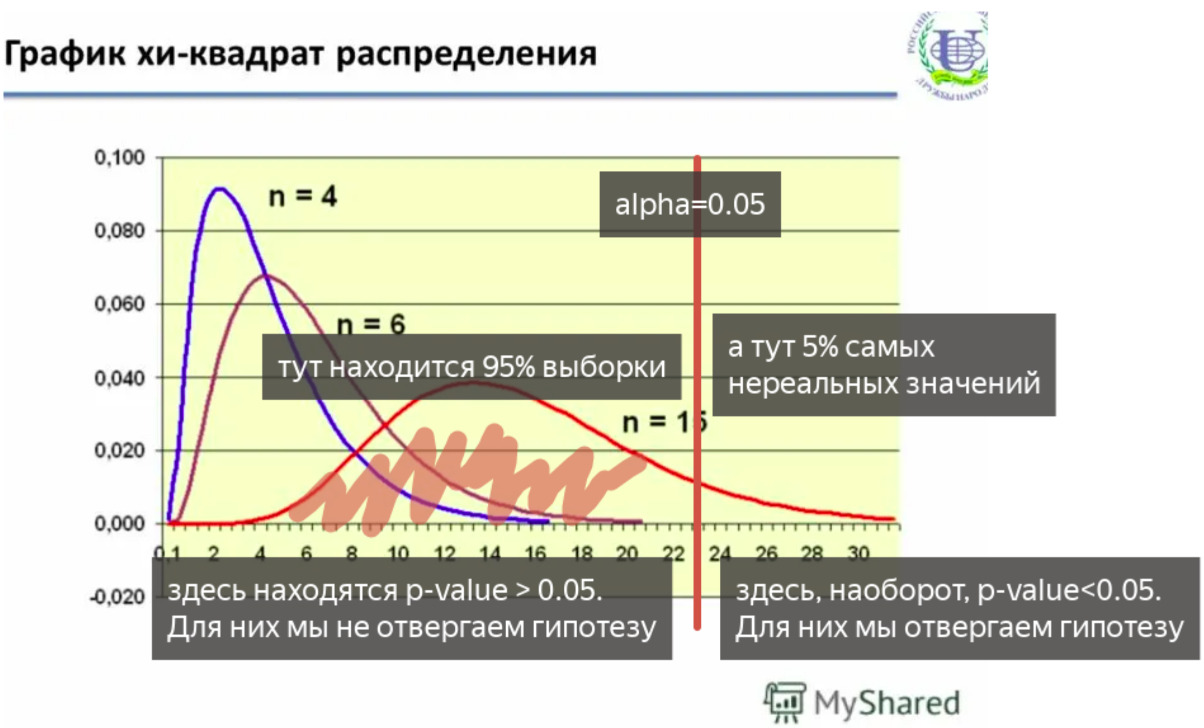
$$\underset{\text{expected}}{E_{ij}} = \frac{\sum_{j=1}^{m} N_{ij}}{N} \cdot \frac{\sum_{i=1}^{n} N_{ij}}{N}$$

Пример

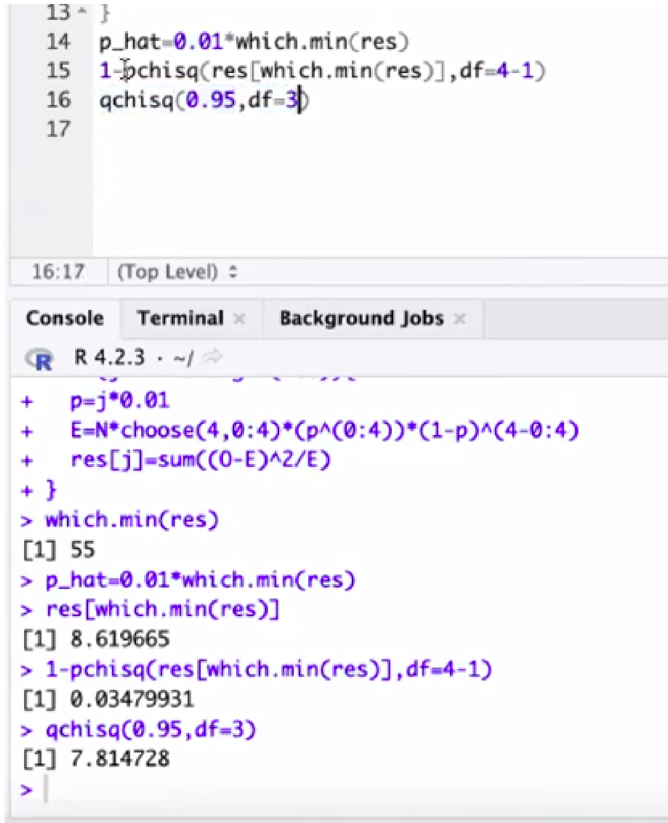


Размышления про p-value

Если p-value меньше уровня значимости (например $\alpha = 0.05,\ p_{value} = 0.012$), то гипотеза отвергается. Иначе, если p-value больше уровня значимости, то гипотеза *не отвергается*, но и не принимается (считаем, что не хватает данных).



Вот снизу пример:



У нас получилось значение 8.61966. Значение, соответствующее p-value=0.05 равно 7.814. Наше значение находится правее порога, поэтому мы отвергаем гипотезу.

Для p-value похожая ситуация. **p-value равно площади с правой стороны (площади правого хвоста)**. У нас она меньше 0.05 $\Rightarrow$ наше значение лежит правее порога, в области нереалистичных значений $\Rightarrow$ гипотеза отвергается.

p-value удобнее, чем просто считать квантиль, т.к. мы сразу получаем уровень значимости.

К слову, еще бывают двухсторонние тесты. В них мы смотрим площадь не только правого хвоста, но и левого.

Выводы

- поговорили про тесты, рандомизированные тесты
- перешли к теореме Уилкса
  - если где-то видим сходимость к кси-квадрату, скорее всего корень лежит в этой теореме
  - поговорили про два применения
    - теорема Пирсона
    - таблица сопряженности